ISCI, Volume *13*

Supplemental Information

A Likely Ancient Genome Duplication in the Speciose Reef-Building Coral Genus, *Acropora*

Yafei Mao and Noriyuki Satoh

Supplemental Figures

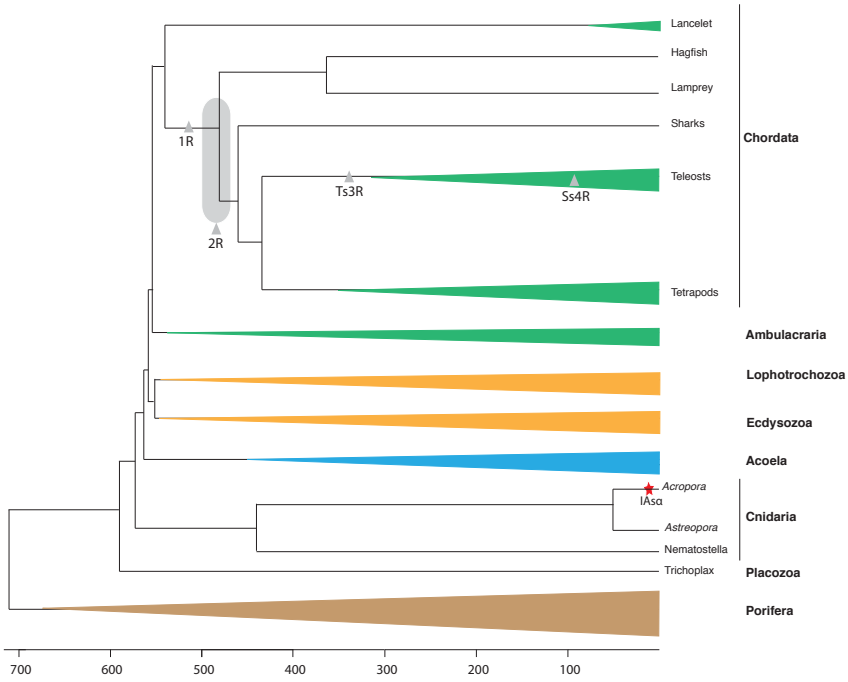


Figure S1. WGD events in evolution of the animal clade. Related to Figure 1. The backbone and divergence time of the tree are based on various sources (e.g., Satoh, 2016). The shaded grey oval represents the uncertain position of two rounds of WGD and colored triangles represent the corresponding divergent groups. Grey triangles represent WGDs and the red star represents invertebrate WGD specific to *Acropora* (IAsα) reported in this study.

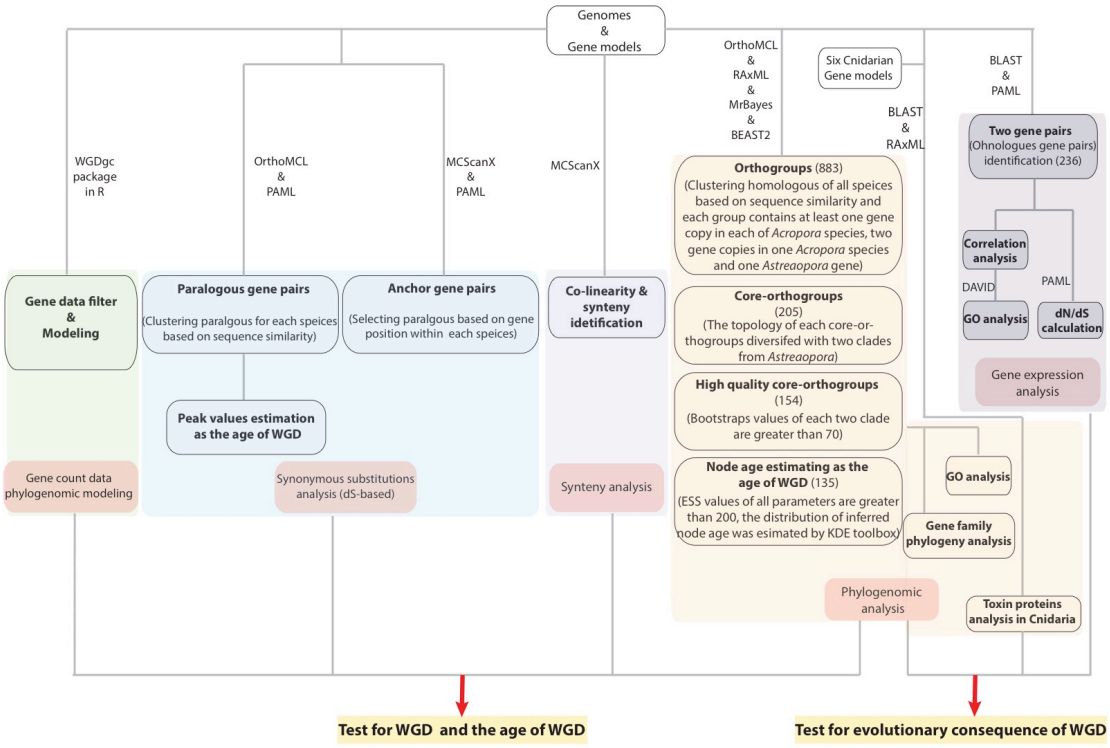
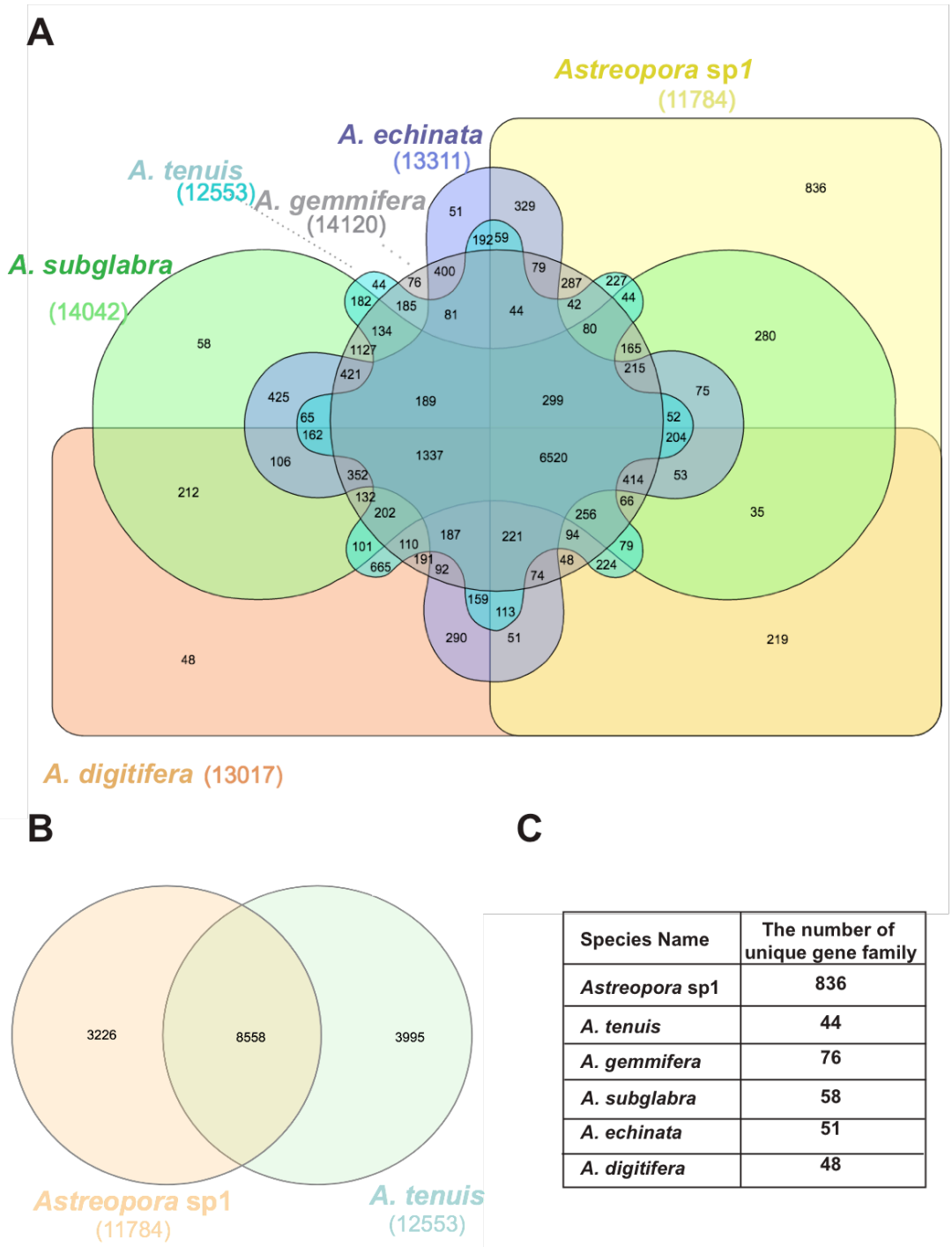


Figure S2. Flowchart of study design. Related to Figures 2-5. The flowchart illustrates methods used to test the main questions in this study.



Species Name	The number of unique gene family
Astreopora sp1	836
A. tenuis	44
A. gemmifera	76
A. subglabra	58
A. echinata	51
A. digitifera	48

Figure S3. Venn diagrams of gene families in six *Acroporid* species. Related to Figure 1. (A). Venn diagram of shared and unique gene families in six *Acroporid* species. (B). Venn diagram of shared and unique gene families between *Astreopora* sp1 and *A. tenuis*. (C). The table of the number of unique gene family in six *Acroporid* species.

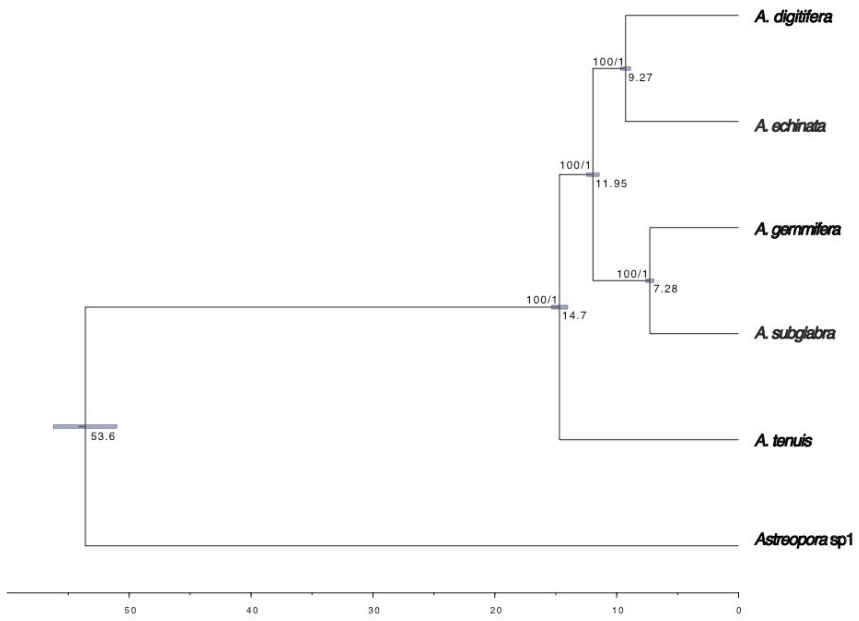


Figure S4. Phylogeny of the Family *Acroporidae.* Related to Figure 1. Time-calibrated phylogenetic tree reconstructed based on fossil calibration and concatenated coding sequences (7,467,066 bp in total) from 3,461 single-copy orthologous genes with BEAST2. Branch lengths are scaled to estimated-divergence time. Posterior 95% CIs of node ages are represented with blue horizontal bars as well as ML bootstrap values and Bayesian posterior probabilities are shown at each node.

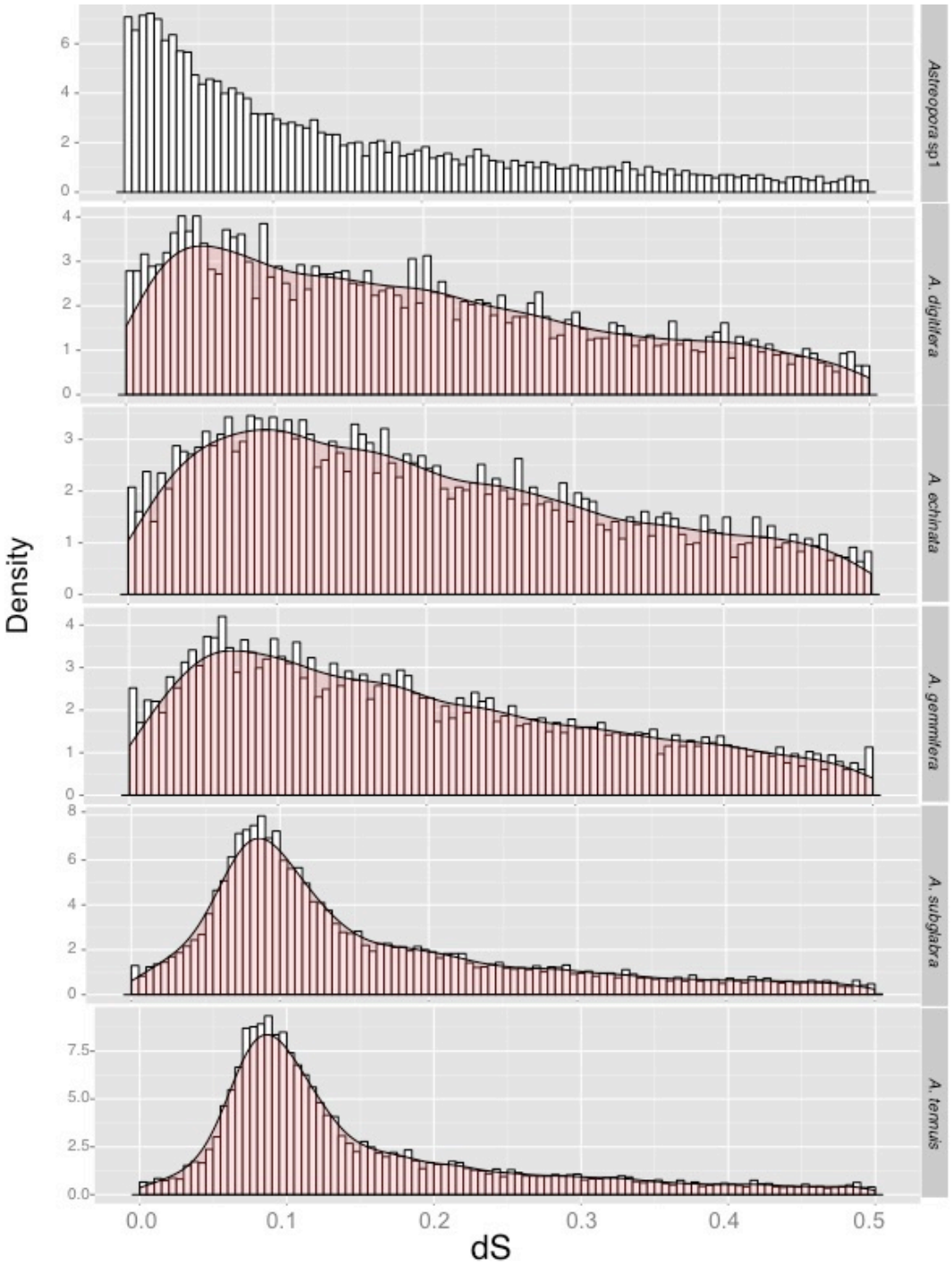


Figure S5. Frequency distributions of dS values for paralogous gene pairs in five *Acropora* and one *Astreopora* species. Related to Figure 2. The distributions of dS values of paralogs, estimating neutral evolutionary divergence since the two paralogs diverged, are plotted with a bin size of 0.005, showing the similar peaks (dS value: 0-0.3) in *Acropora*.

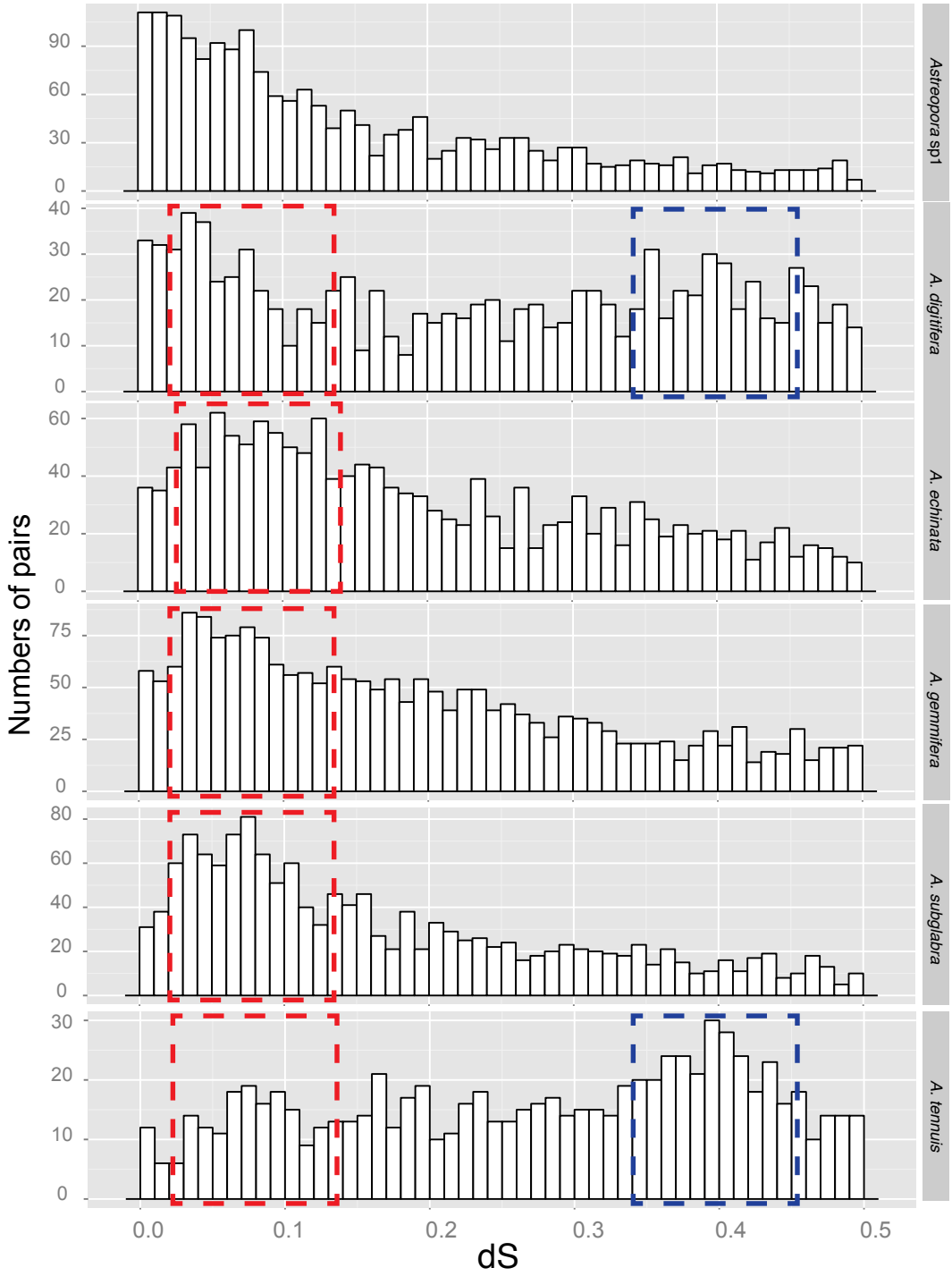


Figure S6. Frequency distributions of dS values for anchor-gene pairs in five *Acropora* and one *Astreopora* species. Related to Figure 2. Distributions of dS values of anchor paralogs, estimating the neutral evolutionary divergence times since the paralogs diverged, are plotted with a bin size of 0.01, showing the similar peaks (dS value: 0-0.3, red boxes) in *Acropora* and extra peaks in *A. digitifera* and *A. tenuis* (dS value: 0.3-0.5, blue boxes).

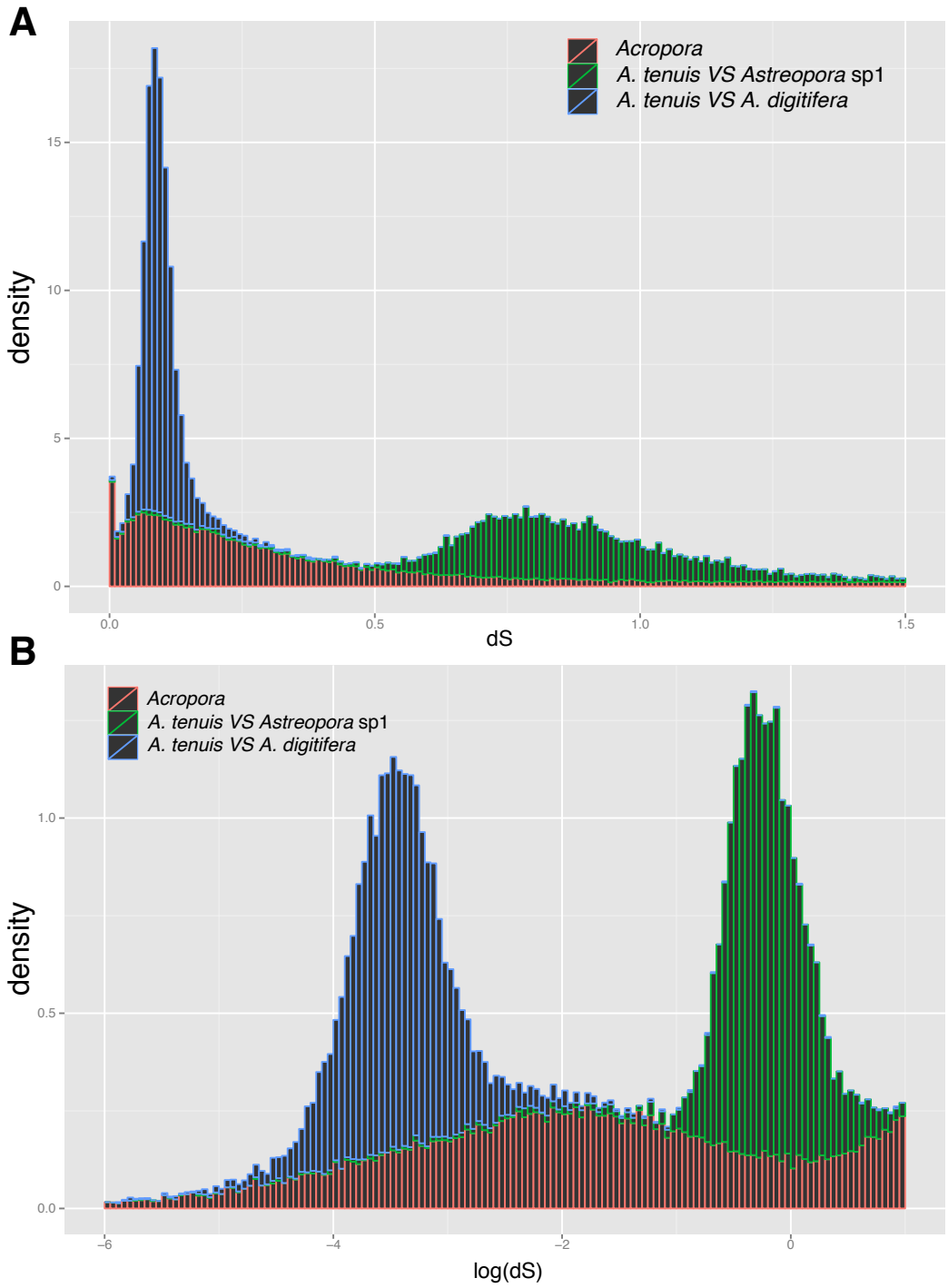


Figure S7. Frequency distributions of dS values for paralogous genes in *Acropora* and for orthologous genes. Related to Figure 2. (A) Frequency distribution of dS values for paralogous genes in *Acropora* and for orthologous genes showing that a WGD event occurred in the most recent common ancestor of *Acropora*. Distributions are plotted with a bin size of 0.01. (B) Frequency distribution of log dS values for paralogous genes in *Acropora* and for orthologous genes. Distributions are plotted with a bin size of 0.05.

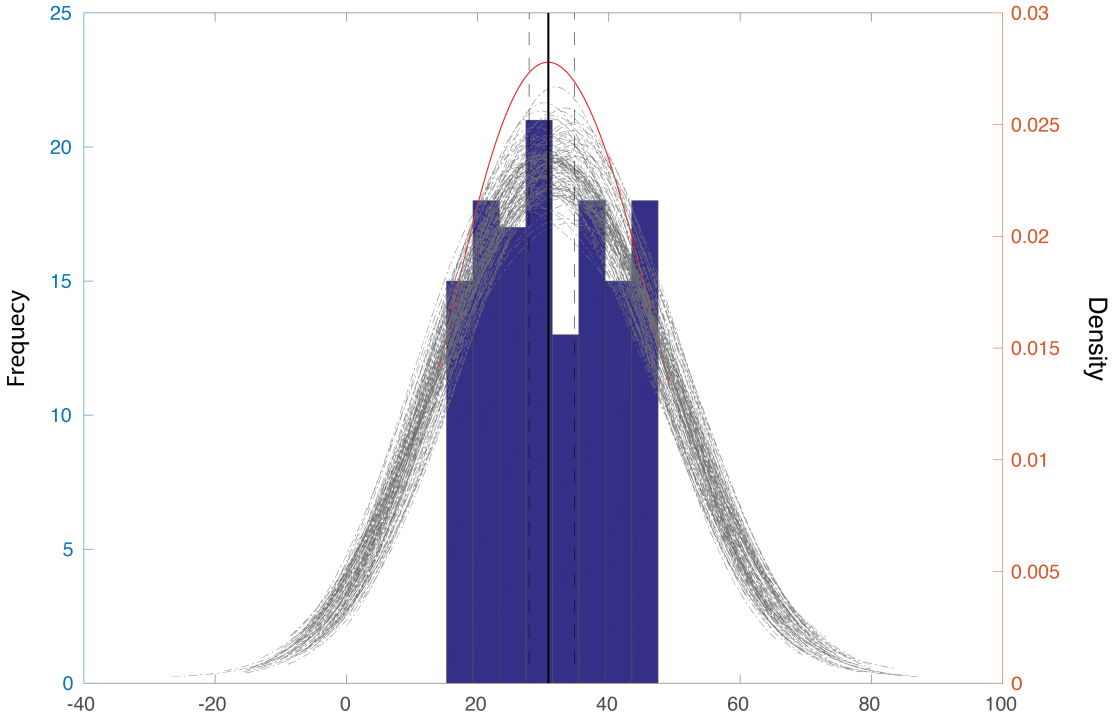


Figure S8. Node age distribution of IAsα. Related to Figure 2. Inferred node ages from 135 phylogenies were analyzed with KDE toolbox to show the peak at 30.78 My, represented by the black solid line. The grey lines represent density estimations from 1000 bootstraps and the black dotted line represents the corresponding 95% confidence interval (27.86 - 34.77 My) from 100 bootstraps.

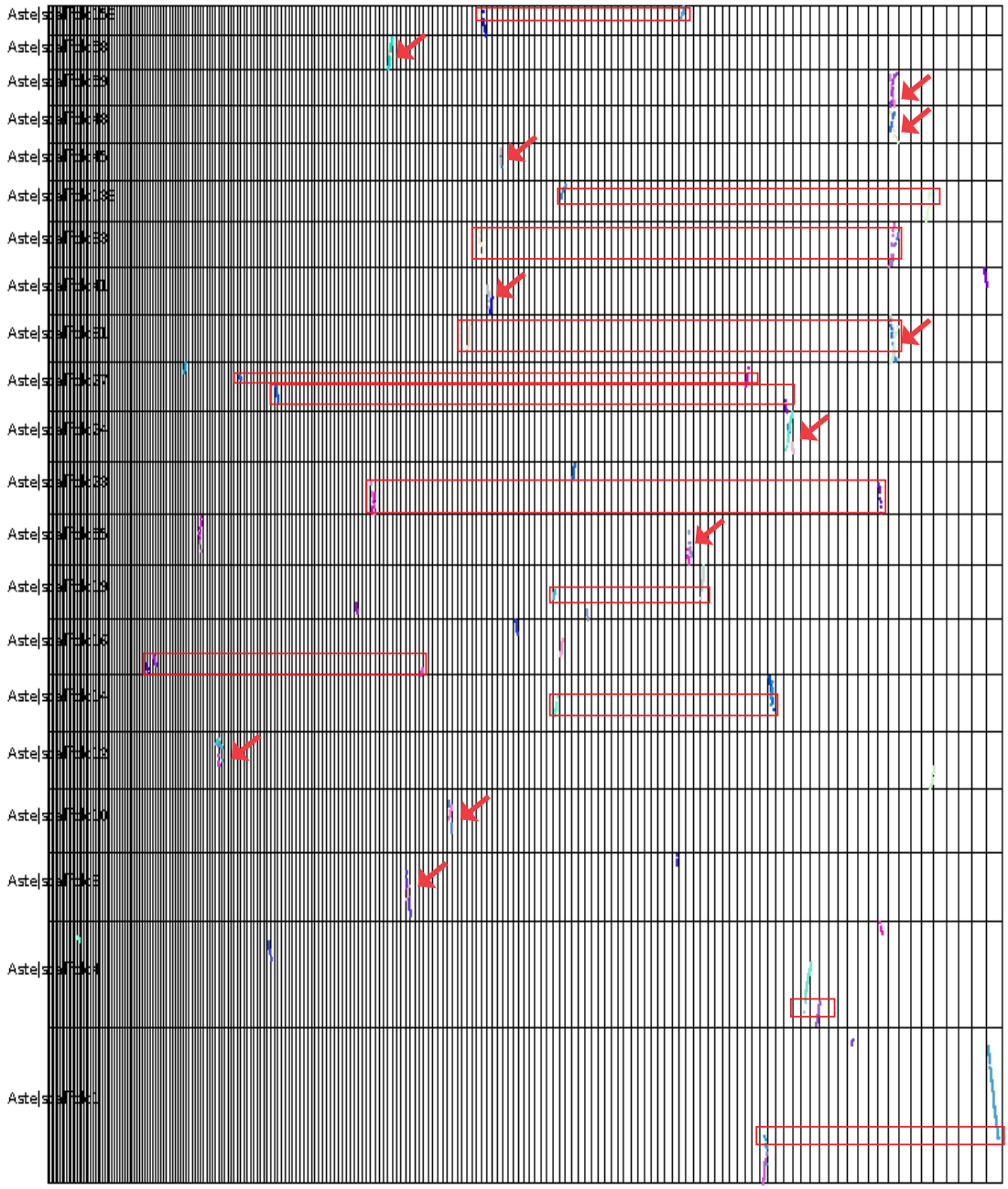


Figure S9. Synteny blocks between *Astreopora* sp1 and *A. tenuis*. Related to Figure 2. Only co-linear segments with at least 8 anchor pairs are shown in between the top length 100 scaffolds of *Astreopora* sp1 (Left side) and the top length 200 scaffolds of *A. tenuis* (Bottom). Only the scaffolds of *Astreopora* sp1 representing duplicated segments with *A. tenuis* are shown. The duplicated segments on different scaffolds are covered with red boxes. The duplicated segments on the same scaffolds are marked with red arrows.

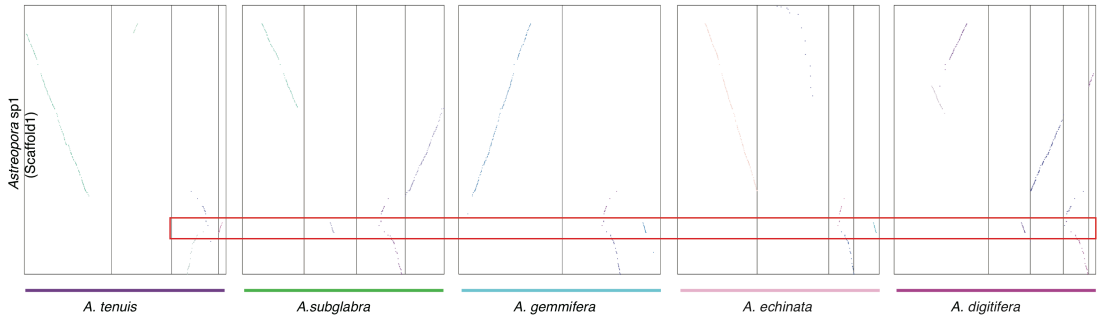


Figure S10. An example of synteny blocks between *Astreopora* sp1 and the five *Acropora* species. Related to Figure 2. Only co-linear segments with at least 8 anchor pairs are shown in between the longest scaffolds of *Astreopora* sp1 (Left side) and scaffolds of other five *Acropora* species (Bottom). The red box represents the retained the duplicated segments.

Figure S11. Alignment of orthogroup 1247 (dnaJ homolog subfamily B member 11-like) showing the independent loss of the domain in duplicates. Related to Figure 3.

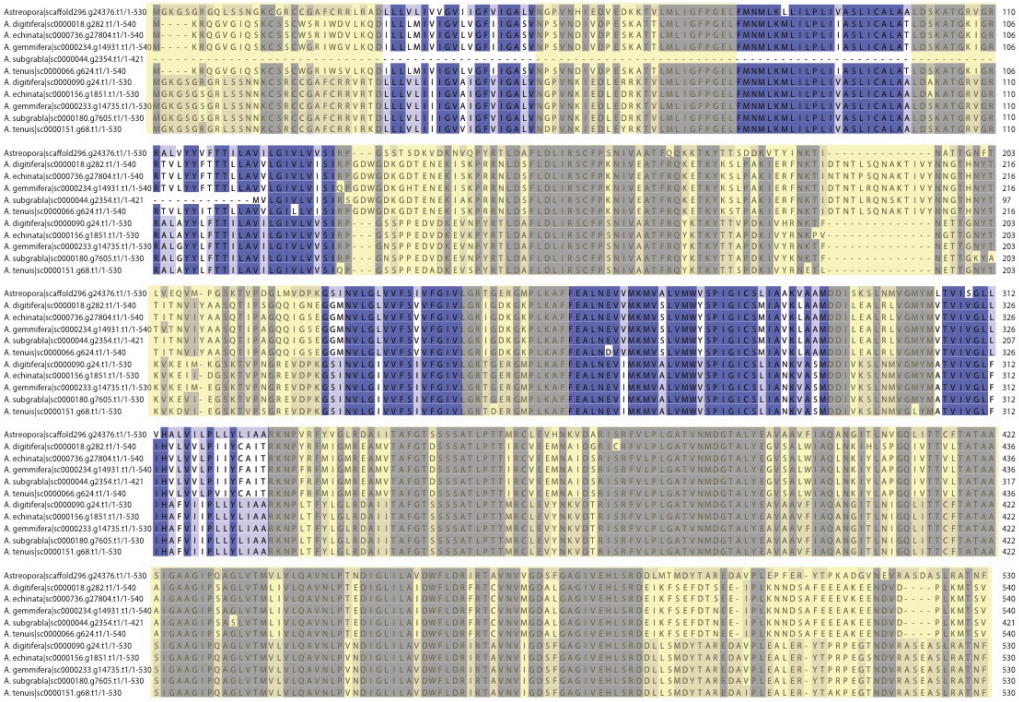


Figure S12. Alignment of orthogroup 1244 (excitatory amino acid transporter 1-like) showing mutations on transmembrane and exposed regions, suggesting that new functions would be generated. **Related to Figure 3.** Exposed regions are shown in yellow.

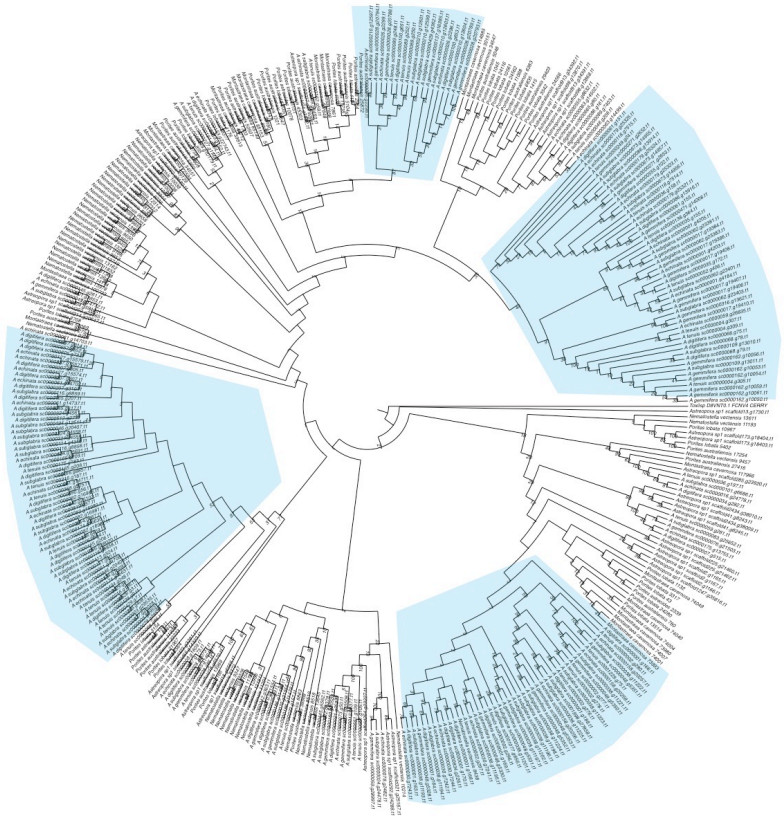


Figure S13. Phylogeny of the toxic protein, Ryncolin-4. Related to Figure 5. The phylogeny was reconstructed using ExaML and bootstrap values are shown at each node. Gene duplications caused by WGD in *Acropora* are shown in cyan.

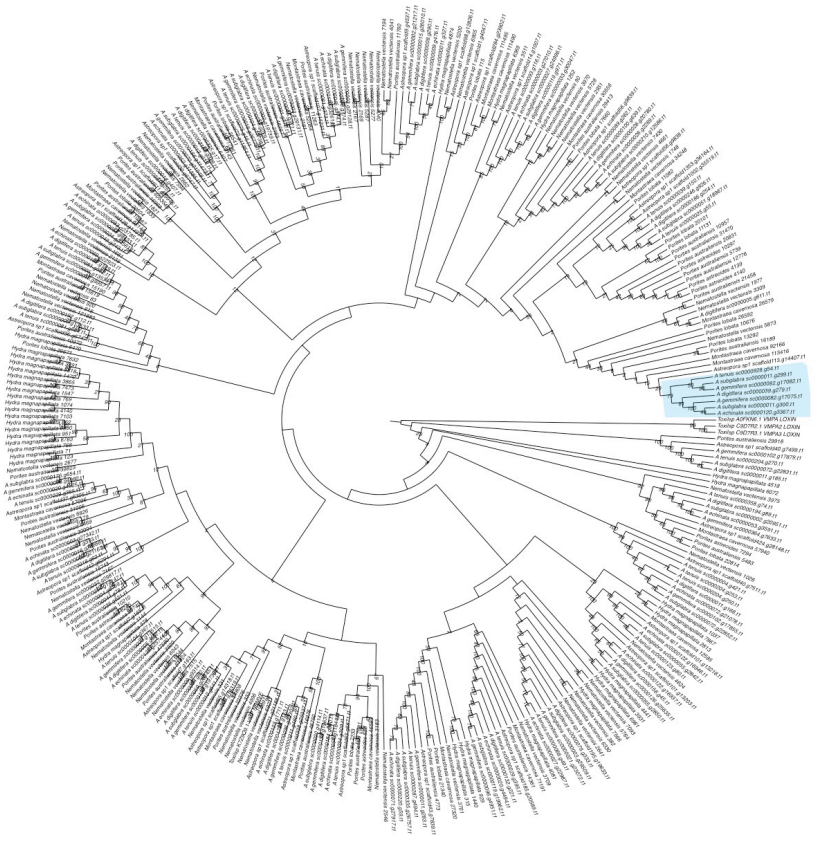


Figure S14. Phylogeny of the toxic Astacin-like metalloprotease. Related to Figure 5. The phylogeny was reconstructed using ExaML and bootstrap values are shown at each node. Gene duplication caused by WGD in *Acropora* is shown in cyan.

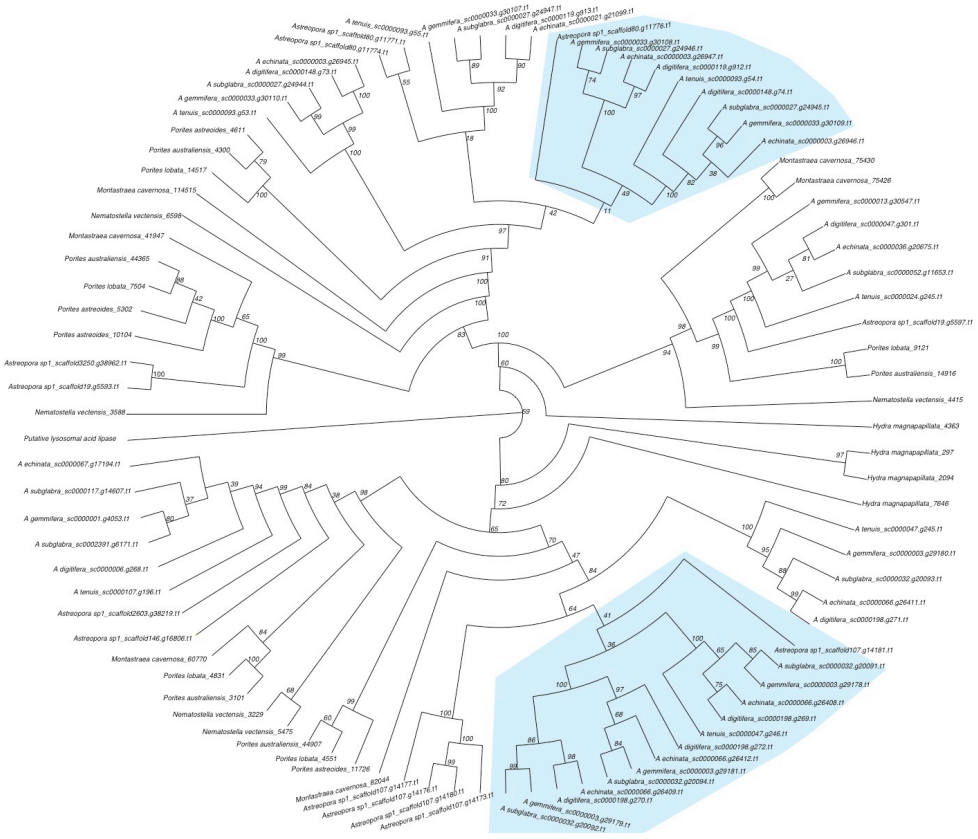


Figure S15. Phylogeny of a toxic protein (Putative lysosomal acid lipase/ cholesteryl ester hydrolase). Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplications caused by WGD in *Acropora* are shown in cyan shadows.

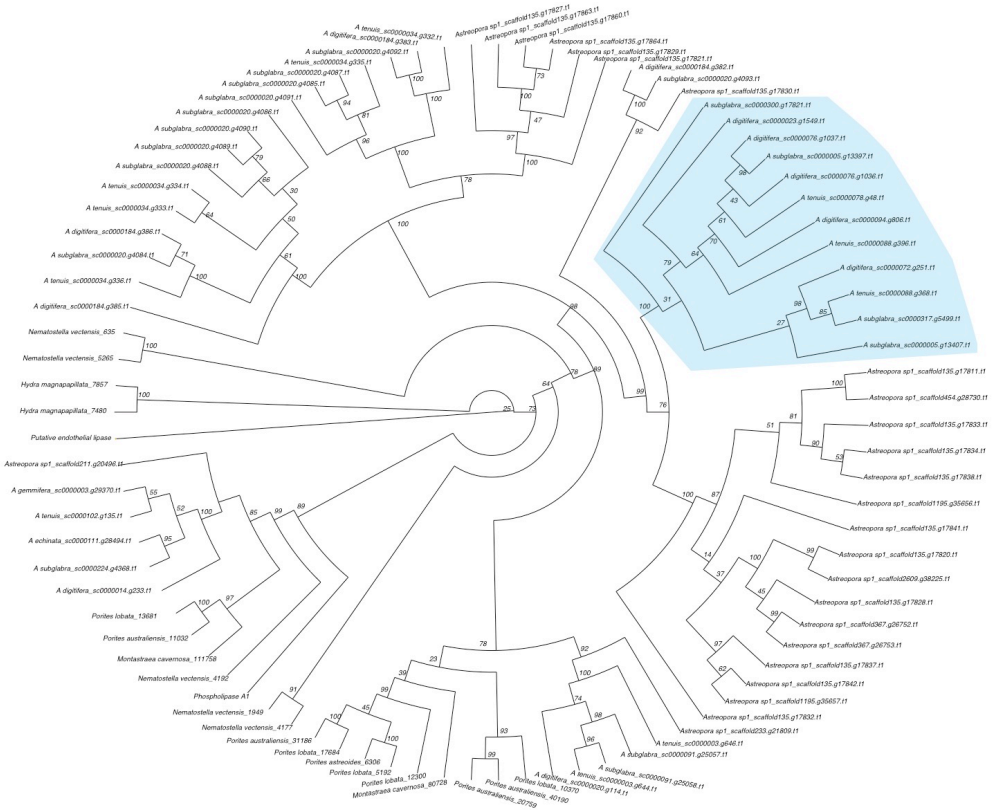


Figure S16. Phylogeny of the toxic putative endothelial lipase. Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplications caused by WGD in *Acropora* is shown in cyan.

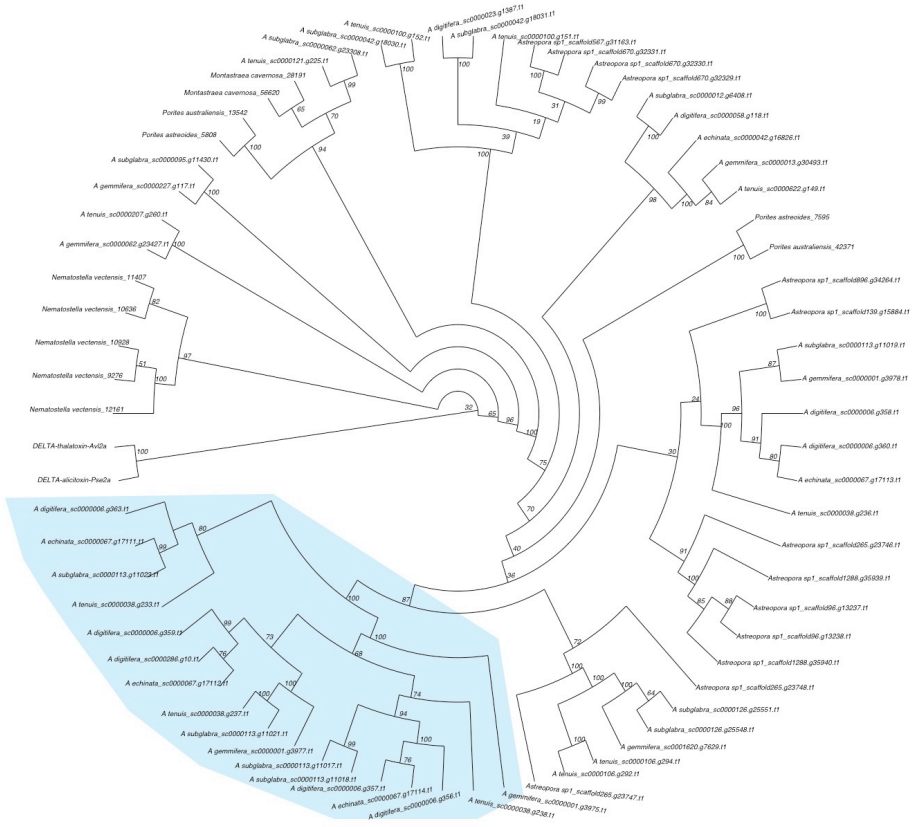


Figure S17. Phylogeny of the toxin protein (DELTA-thalatoxin-Avl2a/DELTA-alicitoxin-Pse2a). Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplication by WGD in *Acropora* is shown in cyan.

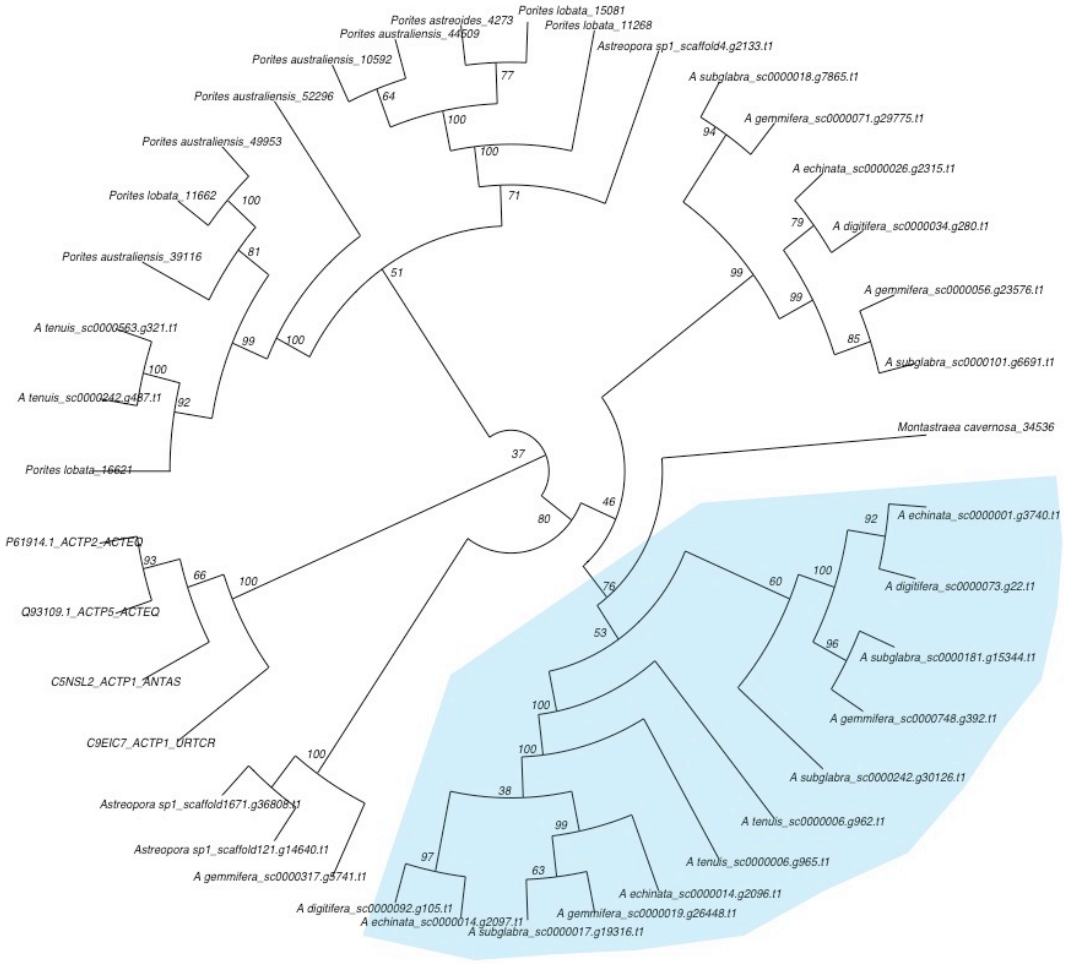


Figure S18. Phylogeny of the toxic protein (DELTA-actitoxin-Aas1a). Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplication by WGD in *Acropora* is shown in cyan.

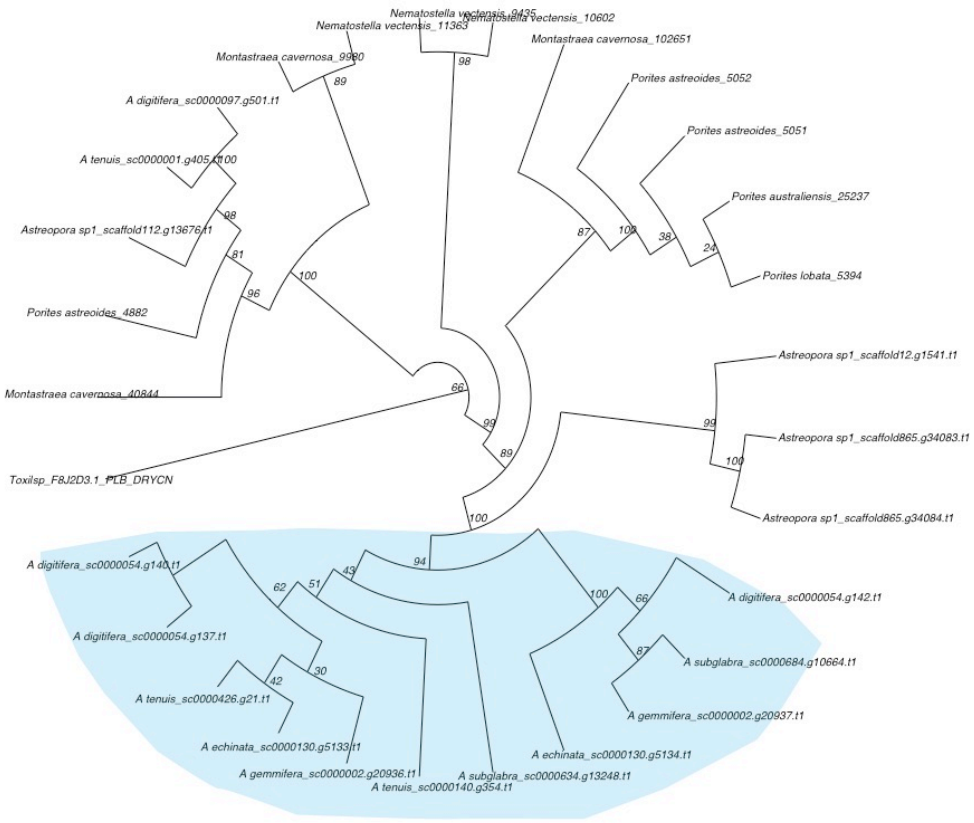


Figure S19. Phylogeny of the toxic protein (Phospholipase-B 81). Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplication by WGD in *Acropora* is shown in cyan.

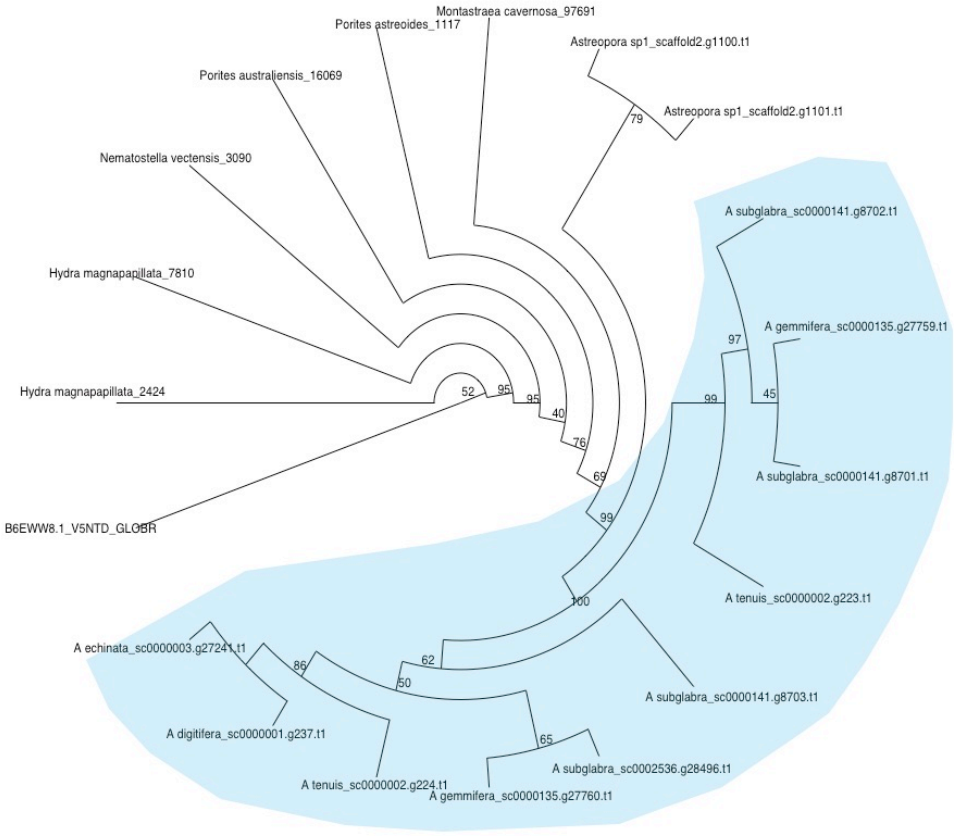


Figure S20. Phylogeny of the toxic snake venom 5'-nucleotidase. Related to Figure 5. The phylogeny was reconstructed using RAxML and bootstrap values are shown at each node. Gene duplication by WGD in *Acropora* is shown in cyan.

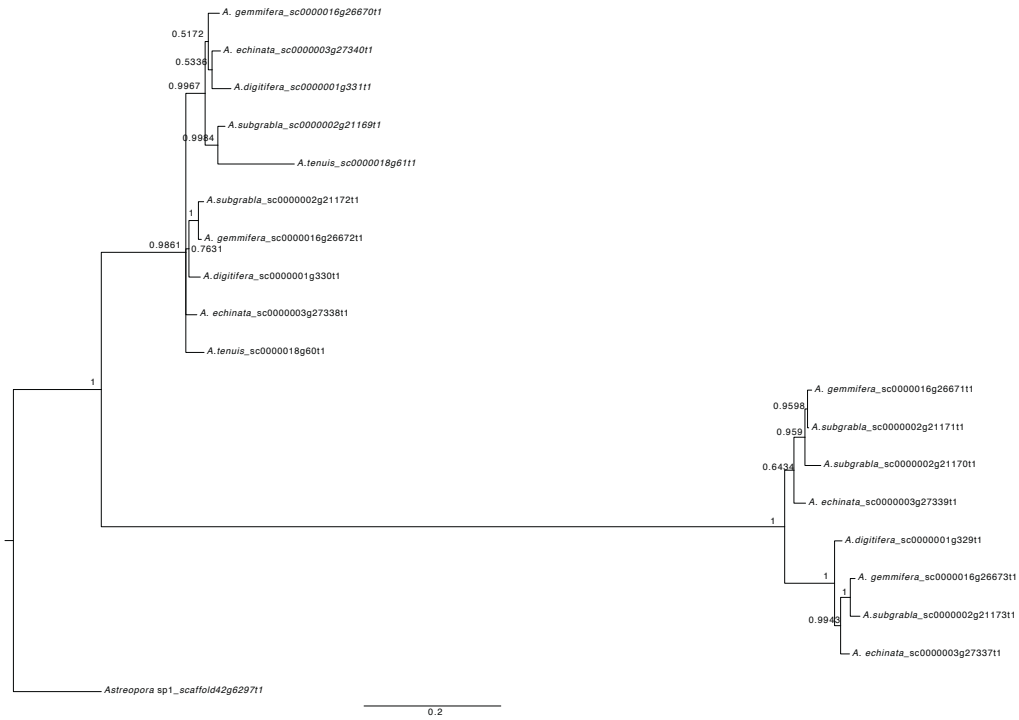


Figure S21. Phylogeny of orthogroup 434 (somatostatin receptor type 5-like) shows duplicates are under two WGD topology. Related to Figure 1. The phylogeny was reconstructed using MrBayes, and Bayesian posterior probabilities are shown at each node.

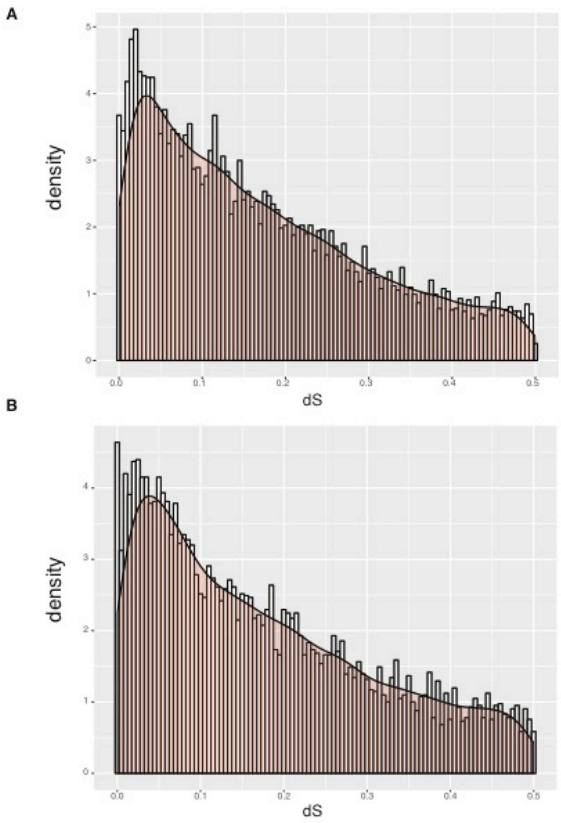


Figure S22. Frequency distributions of dS values for paralogous gene pairs in two new released genomes: (A) *Acropora millepora* and (B) *Acropora tenuis*. Related to Figure 2. The distributions of dS values of paralogs, estimating neutral evolutionary divergence since the two paralogs diverged, are plotted with a bin size of 0.005, showing the similar peaks (dS value: 0-0.3) in *Acropora*.

Supplemental Tables

Table S1. Numbers of gene pairs in the paralogous gene pairs and anchor gene pairs datasets. Related to Figure 2.

	Paralogous gene pairs (<=20 gene families)		Anchor gene pairs (<=20 gene families)	
	Total numbers	Total numbers (0<dS<2)	Total numbers	Total numbers (0<dS<2)
A. digitifera	39827	8249	46559	1958
A.echinata	47299	10948	54956	2530
A. gemmifera	48051	11093	56972	3299
A. subgrabla	50852	12077	44093	2380
A. tenuis	34097	6635	28073	1488
Astreopora sp1	49135	13648	52481	3033

Table S2. Peak value estimations of dS distribution by KDE toolbox. Related to Figure 2.

	Astreopora sp1_*A. tenuis*	*A. tenuis_A. digitifera*	*Acropora_WGD*
Peak age	53.6	14.69	35.01458704
log2(dS_paralog_peak)	-0.314	-3.4596	-1.8165
95%_HDP_log2(dS_paralog_peak)	(-0.22031,-0.33195)	(-3.4008,-3.5141)	(-1.7606,-2.1261)
95%_HDP_Age	NA	NA	(31.18,35.7)

Table S3. Numbers of gene family in orthogroups, core-orthogroups and high-quality core-orthogroups. Related to Figure 2.

	Numbers
Orthogroups	883
Core-orthogroups	205
High-quality core-orthogroups	154

Table S4. The number of putative toxin proteins in 12 Cnidarian species. Related to Figure 5.

Gene family	Query_name	*A. digitifera*	*A.echinata*	*A. gemmifera*	*A. subgla bra*	*A. tenuis*	*Astreo pora sp1*	*H. magni papill ata*	*M. cavern osa*	*N. vecten sis*	*P. astreoi des*	*P. austra liensis*	*P. lobata*
Gene family_1	Coagulation factor X	52	48	52	58	50	49	12	39	56	7	46	34
Gene family_2	Ryncolin-4	48	45	37	62	38	29	0	23	46	5	24	29
Gene family_3	Astacin-like metalloprote ase toxin	28	23	25	26	30	33	36	20	60	6	31	14
Gene family_4	Reticulocalb in	18	15	14	14	18	20	5	16	18	6	19	17
Gene family_5	Putative lysosomal acid lipase/choles teryl ester hydrolase	10	10	10	11	7	13	4	6	5	4	5	5
Gene family_6	Venom carboxyleste rase-6	8	6	9	7	7	17	1	5	14	3	8	4
Gene family_7	Putative endothelial lipase	11	1	1	17	11	25	2	2	5	1	4	5
Gene family_8	DELTA-thalatoxin-Avl2a/DELT A-alicitoxin-Pse2a	9	5	7	12	11	13	0	2	5	2	2	0

Gene family_9	Venom phosphodiesterase 2	5	6	6	6	5	7	1	7	9	3	5	5
Gene family_10	DELTA-actitoxin-Aas1a	3	4	5	5	4	3	0	1	0	1	5	4
Gene family_11		7	5	2	3	2	2	1	1	1	1	2	1
Gene family_12	Phospholipase-B 81	4	2	2	2	3	4	0	3	3	3	1	1
Gene family_13	Venom dipeptidyl peptidase 4	5	2	1	2	2	3	2	4	3	0	0	1
Gene family_14	Hyaluronidase-1	3	2	2	2	2	2	1	2	2	1	2	1
Gene family_15	Snake venom 5'-nucleotidase	1	1	2	4	2	2	2	1	1	1	1	0

Table S5. Likelihood of multiple WGDs hypotheses in *Acropora* using WGDgc method with gene counts data. Related to Figure 1.

WGD event(s)	Likelihood	Likelihood Ratio Test	P_value
0	-38731.86	0 VS 1	7.66E-05
1	-38724.04	1 VS 2	0.01248965
2	-38720.92	2 VS 3	0.05990546
3	-38719.15		

Table S6. Likelihood of different times of WGD under one WGD event in *Acropora* using WGDgc. Related to Figure 1.

Time of WGD	Likelihood
18.697005	-38724.14
22.697005	-38724.09
26.697005	-38724.06
30.697005	-38724.04
34.697005	-38724.04
38.697005	-38724.06
42.697005	-38724.11
46.697005	-38724.2
50.697005	-38724.35

Transparent Methods

Species information, genomic data and gene families cluster

Data can be accessed at: http://marinegenomics.oist.jp, https://przeworskilab.com/data/ and http://comparative.reefgenomics.org/datasets.html (Bhattacharya et al., 2016). The *Acropora* species information in this study was described in our previous paper (Mao et al., 2018). Information about *Astreopora* sp1 was described previously (Suzuki and Nomura, 2013). *Astreopora* sp1 was sampled, sequenced, and assembled in the same way of *Acropora* species (Shinzato et al., in preparation). Transcriptome data of *A. digitifera* across five development stages was described previously (Reyes-Bermudez et al., 2016). Protein sequences of the six species were combined to perform all-against-all BLASTP approach to find all orthologs and paralogs among six species. Then, OrthoMCL was used with default settings to cluster homologs into 19,760 gene families according to sequence similarity (Li et al., 2003).

Single-copy orthologs and reconstruction of a calibrated phylogenomic tree

A custom python script was used to select 3,461 single-copy orthologs with only one gene copy in each species. For each sequence alignment of single-copy orthologs, coding sequences were aligned with MAFFT (Katoh et al., 2002) as described previously (Mao et al., 2018). Then, the concatenated sequences of 3,461 single-copy orthologs were used to reconstruct the phylogenomic tree (species tree) with BEAST2 (Bouckaert et al., 2014). First, we partitioned the concatenated coding sequences by codon position. Molecular clock and trees, except substitution model, were linked together. Then, divergence time was estimated using the HKY substitution model, relaxed lognormal clock model, and calibrated Yule prior with the divergence time estimated in our previous study. We ran BEAST2 three times independently, 50 million Markov chain Monte Carlo (MCMC) generations for each run, then we used Tracer to check the log files and we found that ESS of each of parameters exceeded 200.

Orthogroup selection and detection of a WGD event with dS analysis

(a) dS distributions of paralogous gene pairs

Paralogous gene pairs of each species were identified by all-against-all BLASTP approach and then OrthoMCL was used to cluster paralogs into gene families for each species (Li et al., 2003). Gene families with fewer than 20 genes were used to calculate dS values. Each gene pair within a given gene family was aligned with MAFFT (Katoh et al., 2002) and aligned sequences were used to calculate dS values with Codeml package in PAML with parameters: noisy = 9, verbose = 1, runmode = -2, seqtype = 1, CodonFreq = 2, model = 0, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, and omega = 0.5 (Yang, 2007). The dS distribution of each species was plotted in R (Team, 2013). All processes were run in the GenoDup (Mao, 2019). In addition, in order to avoid bias of genomic data, GenoDup (Mao, 2019) was applied to the two recently released genomes (*Acropora tenuis and Acropora millepora*).

(b) dS distributions of anchor gene pairs

We used MCScanX with default settings (except for match_size=3) to find anchor gene pairs based on synteny information for each species (Wang et al., 2012). Each anchor gene pair was aligned with MAFFT (Katoh et al., 2002) and aligned sequences were used to calculate dS values with Codeml package in PAML with parameters: noisy = 9, verbose = 1, runmode = -2, seqtype = 1, CodonFreq = 2, model

= 0, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, and omega = 0.5 (Yang, 2007). The dS distribution of each species was plotted in R (Team, 2013). All processes were run in the GenoDup (Mao, 2019).

(c) *dS distributions of orthologous gene pairs*

We used MCScanX with default settings (except for match_size=3) to find orthologous gene pairs based on synteny information between *Astreopora* sp1 and *A. tenuis*, and between *A. tenuis* and *A. digitifera* (Wang et al., 2012). Each orthologous gene pair was aligned with MAFFT (Katoh et al., 2002) and aligned sequences were used to calculate dS values with Codeml package in PAML with parameters: noisy = 9, verbose = 1, runmode = -2, seqtype = 1, CodonFreq = 2, model = 0, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, and omega = 0.5 (Yang, 2007). dS distributions of all species were plotted in R (Team, 2013). All processes were run in GenoDup (Mao, 2019).

Detection of a WGD event using phylogenetic analysis

A custom python script was used to select the 883 gene families, including one gene copy in *Astreopora*, one gene copy in each of the five species and at least two ohnologs in one of five *Acropora* species, as orthogroups. Ohnologs are defined as paralogs originating from WGD.

For each of the 883 gene tree reconstructions, we used MAFFT (Katoh et al., 2002) to align amino acid sequences of each single-copy ortholog. We aligned coding sequences with TranslatorX (Abascal et al., 2010) based on amino acid alignments and we excluded the single-copy orthologous genes containing ambiguous 'N'. PartitionFinder (Lanfear et al., 2012) was used to find the best substitution model for RAxML (Version 8.2.2) (Stamatakis, 2014) and MrBayes (Version 3.2.3) (Ronquist et al., 2012), respectively.

Then, 205 orthogroups, for which phylogeny matched the duplication topology (*Astreaopora*, (*Acropora*, *Acropora*)), were selected as core-orthogroups by eyes. The 154 high quality core-orthogroups, for which clades' bootstrap values in ML phylogeny exceeded 70, were used to perform molecular dating with BEAST2 based on the calibrated phylogenomic tree (Bouckaert et al., 2014). Molecular clock and trees, except substitution model, were linked together. Then, divergence time was estimated using the HKY substitution model, relaxed lognormal clock model, and calibrated Yule prior with the divergence time from our previous study. We ran BEAST2 three times independently, 30 million Markov chain Monte Carlo (MCMC) generations for each run. Then we used Tracer to check the log files. 135 time-calibrated phylogenies with ESS values exceeded 200 were selected.

Estimating peak values in dS distributions and inferred node ages' distribution with KDE toolbox

Each distribution was estimated using KDE toolbox in MATLAB, as described previously (Zhang et al., 2017).

(a). Estimating peak values in distributions

To estimate the age of WGD within dS distributions, we assumed the peak value in orthologous gene pair dS distributions as the split time between two species: the split time between *Astreopora* sp1 and *A. tenuis* is 53.6 My, whereas the split time between *A. tenuis* and *A. digitifera* is 14.69 My. Before we used the *kde()* function in KDE toolbox, we first truncated dS distributions to avoid estimation bias due to extreme values: the dS distribution of orthologous gene pairs between *Astreopora* sp1 and *A. tenuis* was truncated with a range from -1 to1 while the dS distribution of

orthologous gene pairs between *A. tenuis* and *A. digitifera* was truncated with a range from -5 to -2. Then, we used the *kde()* function in KDE toolbox to estimate the peak values of these two dS distributions as -0.314 and -3.4596, respectively. Moreover, the distribution of *Acropora* paralogous gene pairs was truncated with a range from -4 to 0 and we estimated the peak value of this distribution as -1.8165. We also used bootstrapping to estimate 95% confidence intervals (CIs) of *Acropora* paralogous gene pairs distribution as -1.7606 to -2.1261 (31.18 to 35.71 My). For bootstrapping, we generated 100 bootstrap samples for each distribution by sampling with replacement from the original data distribution (49,002 samples in the original distribution) with the *sample()* function. We estimated maximum peak values for each 100 bootstrap samples. Then we sorted maximum peak values and values of 6^{th} and 95^{th} rank were used to define the 95% CI.

(b). Estimating peak values in distributions of inferred node age

To estimate the age of WGD in the distribution of inferred node ages, we used the *kde()* function in KDE toolbox to estimate the peak value as 30.78 My, and we used bootstrapping to estimate the 95% CIs as 27.86 to 34.77 My. For bootstrapping, we generated 100 bootstrap samples from the distribution by sampling with replacement from the original data distribution (135 samples in the original distribution) with the *sample()* function. We estimated maximum peak values for each of 100 bootstrap samples. Then, we sorted maximum peak values and values of 6^{th} and 95^{th} rank defined the 95% CI.

Maximum likelihood approach to detect WGD with gene family count data

First, we filtered gene family cluster data generated by OrthoMCL described above (Li et al., 2003). The gene family, including only one *Astreopora* sp1 gene and at least one gene in each of the five *Acropora* species, was counted. Then, we used the WGDgc package in R to estimate log likelihood for parameters (0, 1, 2, 3) of WGD event(s) with setting (dirac=1,conditioning="twoOrMore") (Rabier et al., 2014). Then, we performed likelihood ratio test (pchisq(2*(Likelihood_1-Likelihood_2), df=1, lower.tail=FALSE)) to find the best model and found that one WGD event was the best model to fit the gene family count data. We estimated the age of WGD on 4 My intervals between 18.69 and 38.69 My under a one WGD event model. The lowest log likelihood was shown at the age of WGD: 30.69 and 34.69 My.

Gene expression profiling analysis and dN/dS calculation

We selected 236 gene pairs of *A. digitifera* (ohnologous gene pairs) from 883 orthogroups. We BLASTed these ohnologous gene pairs against the gene expression data across five developmental stages (Reyes-Bermudez et al., 2016) and these data were normalized for each developmental stage. Correlations between two ohnologous genes were performed using Pearson's correlation in R (Team, 2013). Hierarchical clustering was performed using Pheatmap for HC cluster genes and NC cluster genes, respectively. Pairwise dN/dS ratios were calculated with PAML using codeml based on the coding sequence alignment of ohnologous gene pairs with parameters: noisy = 9, verbose = 1, runmode = -2, seqtype = 1, CodonFreq = 2, model = 0, NSsites = 0, icode = 0, fix_kappa = 0, kappa = 1, fix_omega = 0, and omega = 0.5 (Yang, 2007). The dN/dS distribution was plotted with ggplot2 in R and significance tests of differences between dN/dS distributions were evaluated by a Mann-Whitney test in R (Team, 2013).

Evolution analysis of toxic proteins in corals

The 55 toxic proteins of *A. digitifera* identified in the previous study were downloaded from http://www.uniprot.org/ as queries. The protein sequences of *Porites astreoides*, *Porites australiensis*, *Porites lobata*, *Montastraea cavernosa*, *Hydra magnipapillata* and *Nematostella vectensis* were downloaded from http://comparative.reefgenomics.org/datasets.html (Bhattacharya et al., 2016), and combined them with protein sequences of six *Acroporid* species to create a search database.

We identified candidates of toxic proteins by BLASTing the 55 toxins against the combined protein sequences with settings: e-value < $1e^{-20}$ and identity > 30%. Then, we used OrthoMCL to cluster candidates of toxins into 24 gene families and reconstructed their ML gene trees with ExaML (Kozlov et al., 2015) and RAxML. Each gene tree was rooted at a branch or clade of query sequences.

Gene ontology enrichment for duplicated genes of core-orthogroups and protein domains and transmembrane helices prediction

We BLASTed the sequences of 154 high quality core-orthogroups of *Acropora* against the UNIPROT database to find best hits. Identical hits in each ohonlogs group were removed and the remaining hits were used to perform gene enrichment in David (Huang et al., 2009). We also used InterProScan (Zdobnov and Apweiler, 2001) to predict protein domains and used the TMHMM Server (v. 2.0) (Krogh et al., 2001) to predict transmembrane helices from protein sequences.

Supplemental References

Abascal, F., Zardoya, R., and Telford, M.J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic acids research *38*, W7-13.

Bhattacharya, D., Agrawal, S., Aranda, M., Baumgarten, S., Belcaid, M., Drake, J.L., Erwin, D., Foret, S., Gates, R.D., Gruber, D.F., et al. (2016). Comparative genomics explains the evolutionary success of reef-forming corals. Elife *5*.

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology *10*, e1003537.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols *4*, 44-57.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome research *24*, 1384-1395.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research *30*, 3059-3066.

Kozlov, A.M., Aberer, A.J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics *31*, 2577-2579.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology *305*, 567-580.

Lanfear, R., Calcott, B., Ho, S.Y.W., and Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution *29*, 1695-1701.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research *13*, 2178-2189.

Mao, Y., 2019. GenoDup Pipeline: a tool to detect genome duplication using the dS-based method. PeerJ, 7, p.e6303.

McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., dePamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.W. and Leebens-Mack, J.H., (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome biology and evolution, 8(4), pp.1150-1164.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology *61*, 539-542.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

Team, R.C. (2013). R: A language and environment for statistical computing.

Wang, Y.P., Tang, H.B., DeBarry, J.D., Tan, X., Li, J.P., Wang, X.Y., Lee, T.H., Jin, H.Z., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic acids research *40*.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution *24*, 1586-1591.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.